The KH module has an $\alpha\beta$ fold

Maria Antonietta Castiglione Morelli[b], Gunter Stier[a], Toby Gibson[a], Catherine Joseph[a], Giovanna Musco[a,*], Annalisa Pastore[a], Gilles Travè[a]

[a]*EMBL, Meyerhofstr. 1, W-69012 Heidelberg, Germany*
[b]*Università della Basilicata, Potenza, Italy*

Received 28 November 1994; revised version received 14 December 1994

Abstract The KH module has recently been identified in a number of RNA associated proteins including vigilin and FMR1, a protein implicated in the fragile X syndrome. In this work, NMR spectroscopy was used to determine the secondary structure in solution of a KH domain (repeat 5 from vigilin). Almost complete assignments were obtained for the ^{1}H and ^{15}N resonances using uniform ^{15}N-labeling of the protein combined with homo-nuclear 2D ^{1}HNMR and 3D ^{15}N correlated ^{1}H NMR. On the basis of NOE patterns, secondary chemical shifts and amide solvent exposure, the secondary structure consists of an antiparallel three stranded β sheet connected by two helical regions. This domain may also be stabilized by an appended C-terminal helix which is common to many but not all members of the KH family.

Key words: KH; Module; Vigilin; Structure; NMR

1. Introduction

The recently described KH domain (for hnRNP-K homologous domain) [1,2] is one of four motifs (RNP, dsRBD, RGG/RGY and KH) which frequently recur in proteins associated with RNA, often in multiple copies (for a review see [3]). FMR1, which contains two KH domains [4,5] is of particular interest since mutations of the gene cause fragile X syndrome, the most frequent cause of mental retardation in humans, affecting 1 in 1250 males and 30% of female carriers [6–8]. The KH domain is sometimes present in large numbers: 8 in yeast scp160 and 14 in vigilin, the highest number so far [9,10]. Several KH-containing proteins are known to bind to RNA in vitro or were isolated from RNP complexes.

The determination of the 3D structure of the KH motif would greatly benefit biochemical studies on whether and how it binds to RNA and inspire experiments to elucidate the functional role of the proteins containing this domain. A secondary structure prediction based on a multiple alignment has suggested that the KH motif might form an independently folded unit. Sequence periodicity suggests an $\alpha\beta$ structure with a three stranded sheet connected by either one or two α-helices. As a first step towards the elucidation of the 3D structure of the KH module, we have produced genetically engineered fragments containing the KH motif and examined them by NMR spectroscopy. The sequence was selected from vigilin which has a particularly clear subunit structure making it especially suitable for the selection of individual domains and for studying the precise definition of the domain boundaries [10]. Almost full sequential assignment was achieved for one of the fragments, revealing the secondary structure of the KH domain.

2. Materials and methods

2.1. Expression and purification

The DNA sequences corresponding to the KH domain containing fragments of human vigilin [11] in Fig. 1 were PCR-amplified with engineered *Nco*I and *Kpn*I cloning sites respectively on 5′ and 3′ ends and then cloned into the *Nco*I and *Kpn*I sites of a pET9d derivative. These constructs were expressed in *E. coli* strain BL21 (DE3) and the proteins were induced 3 h in a 10 l culture by addition of 0.2 mM IPTG after the culture had reached an OD of 0.8 at 600 nm.

Purification of the construct used for the NMR work was achieved as follows. The pellet was resuspended in 100 ml of 3 mM $MgCl_2$, 0.1 mM PMSF, 10 mM Tris pH 8.5, 0.1 mg/ml DNase. Cells were lysated by French press and centrifuged. The supernatant was loaded on Q-Sepharose and SP-Sepharose columns connected in series and equilibrated with buffer A (10 mM Tris pH 8.5, 0.1 mM PMSF). A washing step of 50 ml buffer A allowed most bacterial proteins to be retained on the Q-Sepharose column, while the KH domain (calculated isoelectric point 9.93) was only retained on the SP-Sepharose. After removing the Q-Sepharose column, the SP-Sepharose column was developed with a NaCl gradient from 0 to 600 mM. The KH domain waseluted at 300 mM NaCl. The fractions containing the domain were concentrated and loaded on a Sephacryl S-200 superfine gel-filtration column equilibrated with 20 mM Tris, pH 8.0, 50 mM NaCl. The protein was finally dialysed against 5 mM KCl, pH 7.5, and concentrated. Uniformly ^{15}N-labeled samples were obtained from bacteria grown in M9 medium with $^{15}NH_4Cl$ as the sole nitrogen source.

2.2. Nuclear magnetic resonance

The NMR measurements were carried out on a BRUKER AMX-600 using 0.5–1 mM samples in 90% H_2O/10% D_2O or in D_2O at pH 7.0–7.5, 5 mM KCl. 2D NMR spectra were acquired between 290 and 308 K in phase sensitive mode (TPPI). Water suppression was achieved either with WATERGATE [12] or, for hetero-nuclear experiments, with the pulse-sequence suggested by Stonehouse et al. [13]. 2D and 3D TOCSY spectra were acquired using the TOWNY composite pulse cycle [14]. Mixing times used were in the 35–65 ms range for the TOWNY and 80–180 ms for the NOESY experiments. 2048 and 512 data points in the acquisition domain and in t_1, respectively, were used for all the 2D experiments. In the 3D experiment, spectral widths were set to 8065, 1946 and 8065 Hz and a matrix of 180, 112, 1024 points in the F1, F2 and F3 dimensions respectively was acquired and successively zero-filled and fourier transformed to a $256 \times 256 \times 256$ points matrix. Since the solubility of the sample after lyophilization decrease drastically, the degree of exposure of the amides was estimated from solvent saturation experiments. Two HSQC were performed in the absence and in the presence of water presaturation using the same receiver gain (the water was suppressed anyway by gradients either according to [12] or [13]). Peaks were integrated by the UXNMR software and they were considered protected when the percentage ratio

*Corresponding author. Fax: (49) (6221) 387 306.

Abbreviations: 1D, 2D and 3D, One-, two- and three-dimensional; HSQC, heteronuclear single quantum coherence; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser enhancement; NOESY, 2D NOE spectroscopy; TPPI, time-proportional phase incrementation; TOCSY, total correlation spectroscopy.

SSDI 0014-5793(94)01422-1

of the values obtained in the presence and in the absence of solvent saturation was higher then 50% normalized on the least affected peak.

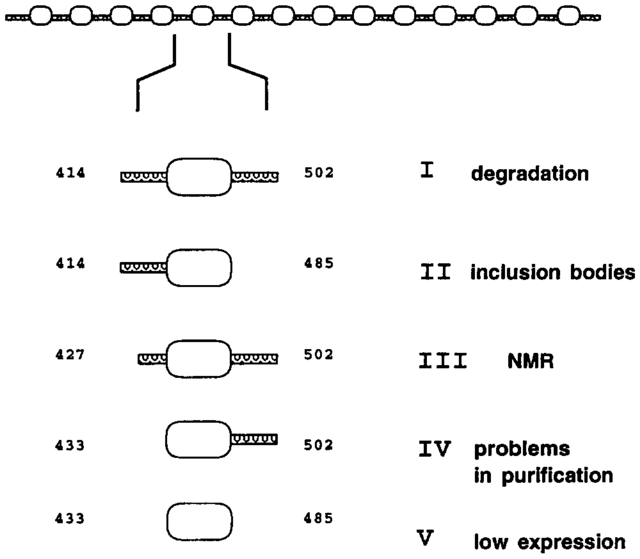
3. Results

3.1. Choice of the domain boundaries

Inspection of known KH-containing sequences showed that in most of them the KH is preceded and followed by a stretch of about 18 residues with strong helical periodicity which links together contiguous KH motifs (see Fig. 1). This pattern is particularly clear for vigilin and the SCP160 protein [9]. We therefore selected one of the vigilin repeats (repeat 5). Of this, five constructs with different boundary choice were expressed but showed different behaviours (Fig. 1). Preliminary 1D NMR experiments showed that constructs I, III and IV are stably folded, the appearance of the spectra being almost superimposable for most of the resonances. Most of the further studies were then carried out on construct III, which was the best behaved.

3.2. NMR assignment

Optimal conditions for solubility and lack of aggregation were found for construct III around pH 7 and at low salt concentration. The protein aggregates at pH below 6, while it unfolds irreversibly at pH 4. The spectra recorded under neutral conditions showed very good intrinsic resolution which significantly helped full spectral assignment. Overlap was also reduced by collecting spectra at different temperature and by a 3D SQC-TOCSY experiment which confirmed the conclusions already obtained by 2D experiments. The assignment was obtained by standard assignment procedures [15]. The identification of two of the three alanines (67 and 75), of the valines



414 427
TEDVNVAQEQIEG MVKDLINRMDYVEINIDHKFHRHLIGKSGAN
485 502
INRIKDQYKVSVRIPPDSEKSNLIRIEGDP *QGVQQAKRELLELAS*

Fig. 1. Schematic representation of the vigilin architecture: ovals indicate the KH motifs, linkers represent putative α-helical regions [10]. The five constructs expressed are indicated below. Numbers on the constructs refer to the corresponding amino acid positions in the complete vigilin sequence from human. For simplicity the numbering adopted in the text refers to the starting of construct III so that for instance Asn-7 corresponds to Asn-434 of the complete sequence. The sequence of construct I is explicitly shown: the sequence of the KH module is indicated in roman letters, while those of the flanking putative helices are in italics.

Table 1
Chemical shifts in ppm of the assigned ^{1}H NMR resonances of the KH module at 600 MHz and 27°C

	^{15}N	HN	CαH	CβH	CγH	CδH	Others
Val-2		8.02	4.00	1.80	0.75		
Lys-3		8.32	4.20				
Asp-4		8.31	4.47	2.62 2.47			
Leu-5		8.15	4.20	1.55		0.81 0.75	
Ile-6		7.97	4.00	1.80	1.10	0.75	
Asn-7		8.40	4.60	2.75 2.62			6.90 7.73
Arg-8	119.87	8.20	4.25	2.20	1.93		
Met-9	119.25	8.27	4.51	2.53 2.39	1.92		
Asp-10	120.75	8.63	4.81	2.62 2.40			
Tyr-11	114.75	7.89	5.51	2.81			6.76 6.61
Val-12	117.00	8.78	4.23	1.78	0.71 0.63		
Glu-13	122.00	8.38	5.25	185 1.78	2.21 1.97		
Ile-14	117.25	8.87	4.52	1.65	0.84 0.75	0.64	
Asn-15	120.75	8.47	5.20	2.63 2.43			7.24 6.77
Ile-16	123.25	8.72	4.20	1.53	1.30 1.00	0.65	
Asp-17	126.80	8.85	4.06	2.56			
His-18		7.60	3.98	3.05			
Lys-19							
Phe-20	115.50	8.27	4.65	3.31 2.91			7.18
His-21	117.90	7.97	4.15	3.13 2.94			7.97 6.37
Arg-22	115.50	8.49	3.99	1.74 1.65			
His-23		7.75	4.41	3.15 3.06			7.70 6.92
Leu-24	117.12	7.80	3.86	1.82	1.29	0.73	
Ile-25	112.62	7.45	3.69	1.81	1.05	0.93	
Gly-26	105.37	7.59	3.90				
Lys-27							
Ser-28	119.62	7.80	4.15	3.75			
Gly-29	106.50	7.80	3.94 3.65				
Ala-30		(7.90)	4.03	1.40			

Table 1 (*continued*)

	^{15}N	HN	CαH	CβH	CγH	CδH	Others
Asn-31	121.50	8.07	4.55	2.83 2.56	7.57 7.02		
Ile-32	117.50	7.68	3.77	1.97	0.86	0.63	
Asn-33	118.75	8.21	4.33	2.78 2.69	7.49 6.82		
Arg-34	119.12	7.54	4.05	2.04	1.84	3.05	
Ile-35	119.00	7.69	3.46	1.87	1.05	0.68	
Lys-36	117.12	8.17	4.01	1.88	1.68 1.50		
Asp-37	116.37	7.75	4.34	2.64 2.54			
Gln-38	116.87	8.54	3.73	1.67 1.09	1.82 1.45		6.87 6.69
Tyr-39	112.00	7.60	4.69	3.23 2.52			7.06 6.59
Lys-40	116.00	7.72	4.10	2.18	182 1.28		
Val-41	107.62	7.31	5.06	1.95	0.54 0.45		
Ser-42	114.00	8.94	4.59	3.67			
Val-43		7.93	4.59	1.71	0.65 0.57		
Arg-44	127.25	9.12	4.48	1.64	1.48 1.23	2.89	
Ile-45	128.12	8.50	4.37	1.70	1.31 0.54	0.75	
Pro-46	–	–	4.05	2.22	2.06	3.53 3.99	
Pro-47	–	–	4.48	2.45	2.00 1.91	3.60 3.70	
Asp-48		(8.20)	(4.48)	2.60 2.48			
Ser-49			4.31	4.02 3.85			
Glu-50		(8.47)	4.20	2.25 1.61			
Lys-51	119.75	(7.59)	(4.00)	(2.19 2.15)	2.00		
Ser-52	112.00	8.45	4.50	3.73 3.46			
Asn-53	120.87	8.43	4.97	2.95 2.63	7.46 6.90		
Leu-54	119.50	7.69	4.61	1.56	1.34	0.73 0.65	
Ile-55	124.12	8.84	4.31	1.53	1.32 0.82 0.45	0.24	
Arg-56	124.37	7.58	5.14	1.62	1.49 1.35	2.87 2.98	
Ile-57	125.75	8.94	4.91	1.59	1.39 0.54 0.75	0.63	
Glu-58	122.50	8.50	5.34	1.96 1.86	2.13		
Gly-59	108.75	8.35	4.46 3.92				
Asp-60	122.50	8.15	4.78	2.83 2.62			
Pro-61	–	–	4.41	1.85	1.05	3.73 3.95	
Gln-62	115.12	8.17	4.05	2.04	2.30 (1.88)		
Gly-63	109.62	8.45	4.03 3.73				
Val-64	118.12	8.60	3.22	1.86	0.62 0.54		
Gln-65	114.37	7.12	3.76	1.95	2.03 2.37		7.64 6.68
Gln-66	117.75	7.47	3.73	1.29	2.21 2.12		7.35 6.80
Ala-67	119.75	8.25	3.46	1.07			
Lys-68	115.75	8.00	3.41	1.79 1.07	1.59 1.37	2.81 2.73	
Arg-69	115.37	7.12	3.78	1.82 1.75	1.60 1.41	3.07	
Glu-70	116.12	7.67	3.88	2.12 2.05	2.41		
Leu-71	117.87	8.32	3.85	2.63 1.35	1.47	0.85 0.42	
Leu-72	114.87	7.84	3.87	1.76	1.29	0.74 0.62	
Glu-73	117.37	7.66	3.92	2.12 1.99	2.35		
Leu-74	118.00	7.59	4.09	1.84	1.38	0.77	
Ala-75	119.87	7.25	3.96	1.08			
Ser-76	118.25	7.18	3.97	3.69			

Tentative assignments are given in parentheses. ^{1}H spectra at 300K were referred to the water resonance at 4.69 ppm.

and of the AMX systems was achieved immediately from 2D TOCSY patterns. The aromatic rings of the two tyrosines (Tyr-11 and -39) could be easily connected to their amide protons by NOESY connectivities. The resonances of the aromatic protons of Phe-20 were equivalent and could be identified only after examination of spectra in D_2O and from an ^{15}N-SQC experiment. Pro-46, Pro-47 and Pro-61 were identified as *trans* residues from the presence of typical NOESY peaks between their δ protons and the α protons of Ile-45, Pro-46 and Asp-60 respectively [15]. In the fingerprint region below 4.5 ppm, strong sequential peaks connected with well isolated peaks which could then be identified easily (Fig. 2). Sequential HN–HN connectivities could be followed in the amide region with almost no overlap. Several independent entrance points helped in placing these connectivities along the sequence.

Both hetero and homo-nuclear spectra showed, in addition to well defined and dispersed peaks, a few broad resonances with poor transfer. Some of these resonances could be assigned to the flexible N-terminus (from Val-2 to Asn-7) by comparing the two N-deleted constructs which differed by the presence of the first six residues (construct III and IV). Six residues (Lys-27, Ala-30, Asp-48, Ser-49, Glu-50 and Lys-51) could not be identified from the amide region both in TOCSY and in NOESY experiments, even at very long mixing times. Since they are in unstructured loop regions, their amide protons are subjected to significant loss of sensitivity, which is especially high at neutral pH. This effect could be reduced but not cancelled using pulse field gradient techniques [12,13]. Ala-30, Ser-49 and Glu-50 could however be observed from their α protons in the TOCSY and tentatively assigned. The spectral assignment is summarized in Table 1.

3.3. *Identification of the secondary structure*

Several resonances below 4.7 ppm and a number of sequen-

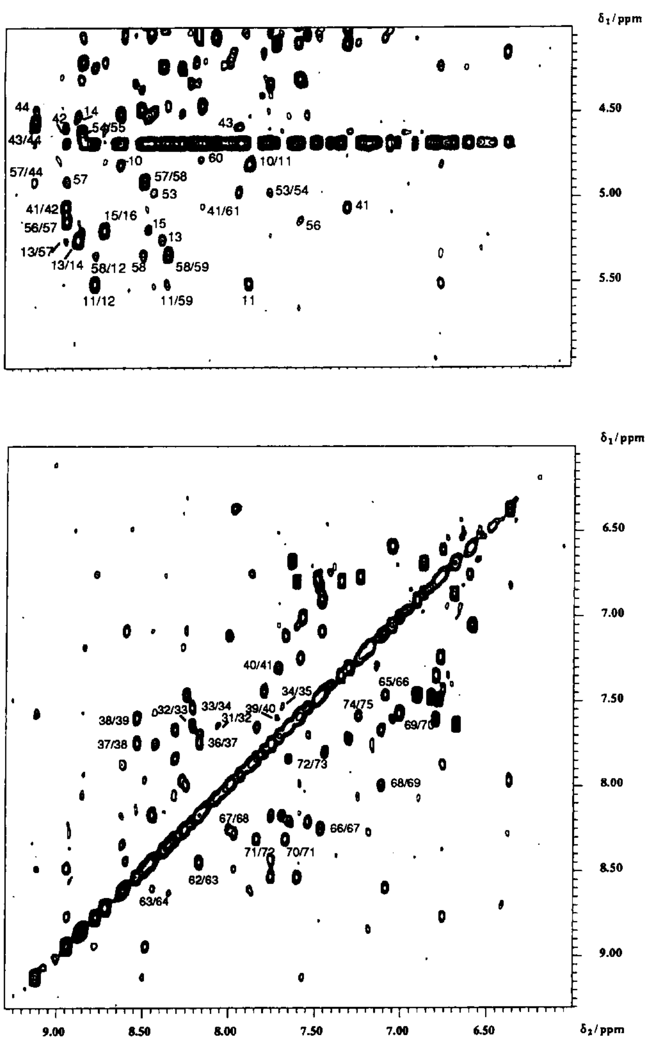


Fig. 2. NOESY spectrum at 300K and 600 MHz of a 1 mM solution sample of repeat 5 of vigilin at pH 7.0. Both the amide/amide and the fingerprint regions are shown. The numbers indicated correspond to thesequential amide/amide connectivities of H2 and H3 (lower plot) and to long-range HA/HN connectivities in the β-sheet.

tial amide/amide NOESY connectivities suggest immediately the presence of $\alpha\beta$ structure. Strong sequential Hα,HN(i,i + 1) NOESY connectivities as well as long-range backbone connectivities (see Fig. 3a and b) showed the presence of a three-stranded β sheet involving residues Met-9 to Ile-16, Tyr-39 to Leu-45 and Asn-53 to Asp-60. The arrangement of the three strands is antiparallel with the first and the second strands flanking the last one. In addition, three helical stretches are present as shown by sequential HN-HN(i,i + 1), HA-HN(i,i + 3) and (i,i + 4) connectivities: a very short and flexible helix-like region between residue His-21 and Gly-26 (designated H1) and a well defined helix spanning Asn-31 to Gln-38 (designated H2) which is connected to the third β strand by a type II β turn. The last helix extends from Gln-62 to Leu-74 and spans the C-terminal linker (designated H3) (Fig. 3c). The seven N-terminal residues and the region between Lys-19 and His-23 are flexible in solution, as shown by the significantly reduced NOE intensities. The C-terminal region is, by contrast, very well structured until the end. The protection of amide protons measured by solvent saturation transfer experiments shows a very good correlation with the secondary structure elements (Fig. 3a). In particular, amides in the region 54–60 are generally more protected than those in the external β-strands. Finally, a plot of the secondary chemical shifts vs. the sequence shows three negative and three positive regions fully consistent with the presence of the three helices and strands (Fig. 3a).

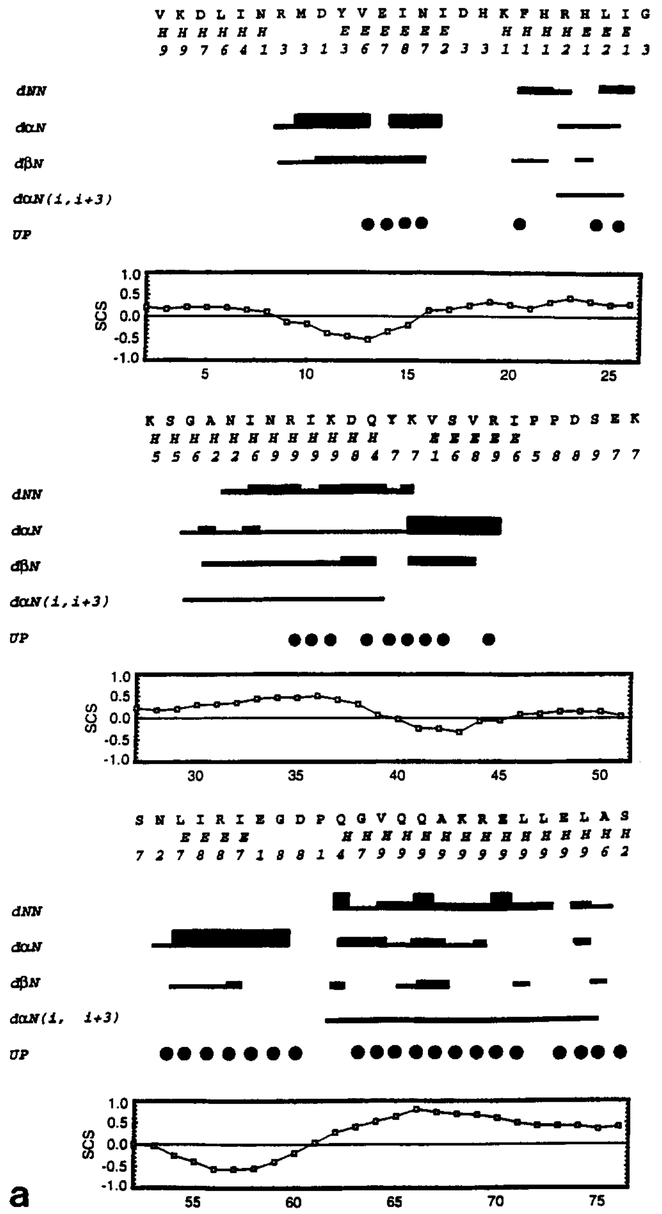


Fig. 3. (a) Survey of the NOE effects observed. In italics is the secondary structure prediction as obtained from the PredictProtein EMBL e-mail server [23] is reported together with the degree of confidence. The main classes of short range contacts observed are shown together with the secondary chemical shifts (SCS) and protons which are solvent inaccessible in solvent saturation transfer experiments (UP). A window of ±2 points was used to smooth the SCS data. (b) The three-stranded β-sheet of the KH. Continuous lines identify observed interstrand NOEs. A dotted line between Ile-14 and Ile-55 indicates that this effect could not be observed because these two amides are coincident. (c) Schematic representation of the secondary structure of the KH domain in the same orientation as in b.

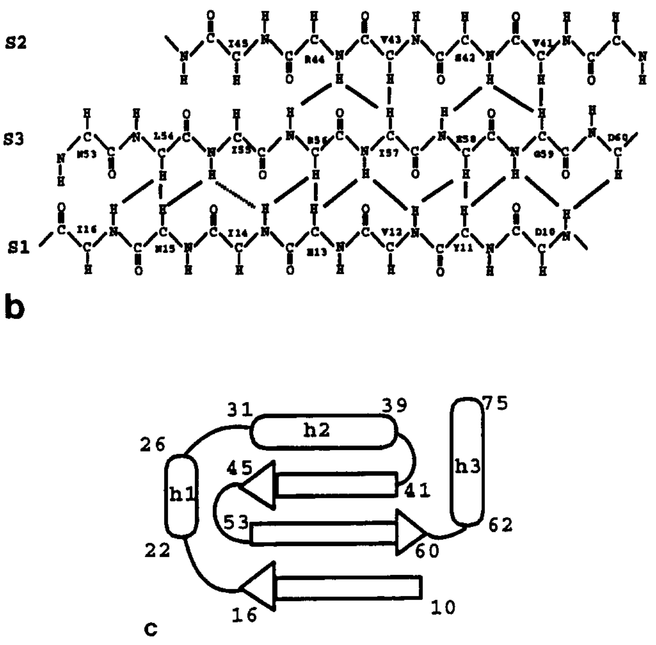


Fig. 3 (*continued*).

3.4. Three-dimensional arrangement

In addition to the long-range effects which identify the presence of a β-sheet, other connectivities were observed even at an early stage of the structure determination. The aromatic ring of His-21 is surrounded by Ile-55 (the δ protons of which are ring current shifted to 0.22 ppm), Pro-46 and Asn-43. The aliphatic protons of Ile-45 and Ile-55 also show clear connectivities. These effects help in placing the H1 approximately perpendicularly to the β-sheet (Fig. 3c). The presence of two conserved hydrophobic residues in H2 (Ile-32 and -35) suggests that they pack into the β sheet but no NOEs were observed so far because of overlap. Preliminary analysis of the NOESY data suggests that in our construct the C-terminal helix folds back and packs into the KH. However, the role of these helices remains perplexing since a survey of the sequence databases shows that all the proteins containing helix–KH–helix repeats contain $n + 1$ helices if n is the number of KH domains. In addition to vigilin (where there are 15 helices but only 14 KH domains) and yeast Scp160 (with 9 helices but 8 KH domains), the family includes hnRNP K [3]; hnRNP E1 (accession no. X78137); Nova-1 [16]; FBP [17]; and three *C. elegans* cosmid reading frames, C06G4.1, M88.5 and Zk418.9 [18]. However, each KH domain appears most closely associated with the C-terminal helix: there is almost never an insertion in the loop connecting the third KH strand and the following helix, whereas large insertions are tolerated between the preceding N-terminal helix and the first β-strand of the KH domain.

4. Discussion

The work described here shows that the KH motif folds into a stable $\alpha\beta$ structure consisting of a three-stranded sheet and two helices, the first of which has high flexibility, in excellent agreement with secondary structure predictions ([2] and Fig. 3a). The KH domain sequence alignment shows no absolutely conserved residues, although there is a preference for positively charged residues in the region encompassing the two KH helices and the connecting loop. Thus these residues will cluster to one face of the KH domain, suggesting a surface for interaction with RNA. The flexible H1 has the highest density of semi-conserved charge (e.g. five in vigilin repeat-5). In some other nucleic acid binding domains, such as the bZIP proteins [19] and the U1A RNP domain [20], flexible regions in the unbound domain acquire a stable fold induced by binding nucleic acid. By analogy with such proteins, we suggest that the behaviour in solution of the region around H1 will be lost upon RNA-binding.

Most fragile X lesions are caused by amplified triplet repeats which completely shut down FMR1 gene expression. However, in one documented case with an aggravated phenotype, the FMR1 protein is expressed but carries a point mutation, Ile-304→Asn, in the second KH domain at a conserved hydrophobic position [21,22]. This amino acid, equivalent to Ile-32 in construct III, is in the second helix of the KH and shows a very high protection factor from solvent saturation transfer experiments. Therefore, the mutation of this residue is likely to destabilize the KH domain structure.

Acknowledgements: We would like to thank prof. Dobberstein for the vigilin clone and for helpful discussion and M. Pfuhl for help in processing 3D data.

References

[1] Siomi, H., Matunis, M.J., Michael, W.M. and Dreyfuss, G. (1993) Nucleic Acids Res. 21, 1193–1198.
[2] Gibson, T.J., Thompson, J.D. and Heringa, J. (1993) FEBS Lett. 324, 361–366.
[3] Burd, C.G. and Dreyfuss, G. (1994) Science 265, 615–621.
[4] Siomi, H., Siomi, M.C., Nussbaum, R.L. and Dreyfuss, G. (1993) Cell 74, 291–298.
[5] Gibson, T.J., Rice, P.M., Thompson, J.D. and Heringa, J. (1993) Trends Biochem. Sci. 18, 331–333.
[6] Mandel, J.L. and Heitz, D. (1992) Curr. Op. Genet. Dev. 2, 422–430.
[7] Webb, T.P., Bundey, S.E., Thake, A.I. and Todd, J. (1986) Am. J. Med. Genet. 23, 573–580.
[8] Gustavson, K.H., Blomquist, H. and Holmgren, G. (1986) Am. J. Med. Genet. 23, 581–588.
[9] Wintersberger, U., Kuchne, C. and Karwan, A., EMBL Database Accession no. X65645).
[10] Schmidt, C., Henkel, B., Poeschl, E., Zorbas, H., Puschke, W.E., Gloe, T.R. and Mueller, P.K. (1992) Eur. J. Biochem. 206, 625–634.
[11] McKnight, G.L., Reasoner, J., Gilbert, T., Sundquist, K.O., Hokland, B.,McKernan, P. A., Champagne, J., Johnson, C.J., Bailey, Mason, C., Holly,R., O'Hara, P.J. and Oram, J.F. (1992) J. Biol. Chem. 267, 12131–12141.
[12] Piotto, M., Saudek, V. and Sklenar, V. (1992) J. Biomol. NMR 2, 661–664.
[13] Stonehouse, J., Shaw, G.L., Keeler, J. and Laue, E.D. (1994) J. Magn. Res. A 107, 178–184.
[14] Kadkhodaei, M., Hwang, T.L., Tang, J. and Shaka, A.J. (1993) J. Magn. Res. 105, 104–107.
[15] Wuethrich, K. (1986) NMR of Proteins and Nucleic Acids, Wiley, New York.
[16] Buckanovitch, R.J., Posner, J.B. and Darnell, R.B. (1993) Neuron 11, 657–672.
[17] Duncan, R., Bazar, L., Michelotti, G., Tomonaga, T., Krutzsch, H., Avigan, M. and Levens, D. (1994) Genes Dev. 8, 465–480.

[18] Wilson et al. (1994) Nature 368, 32–38.

[19] Saudek, V., Pastore, A., Castiglione Morelli, M.A., Frank, R., Gausepohl, H., Gibson, T.J., Weih, F. and Roesch, P. (1990) Protein Engineering 4, 3–10.

[20] Oubridge, C., Ito, N., Evans, P. R., Teo, C.-H. and Nagai, K. (1994) Nature 372, 432–438.

[21] De Boulle, K., Verkerk, A.J.M.H., Reyniers, E., Vits, L., Hendrickx, J., Van Roy, B., Van den Bos, F., De Graaff, E., Oostra, B.A. and Willems, P.J. (1993) Nature Genet. 3, 31–35.

[22] Siomi, H., Choi, M.C., Nussbaum, R.L. and Dreyfuss, G. (1994) Cell77, 33–39.

[23] Rost, B. and Sander, C. (1994) Proteins 19, 55–72.